Predicting Alzheimer's Disease

An End-to-End Supervised Learning Workflow and In-depth EDA

By Nour Hatem

Problem & Solution

Global Crisis

Alzheimer's disease is an escalating global health crisis, impacting millions and straining healthcare systems worldwide.

Diagnostic Challenge

Early, accurate diagnosis is crucial for patient management and future planning yet remains a significant clinical challenge.

Our Mission

Leveraging a rich clinical dataset to build, evaluate, and identify the most robust machine learning model for early-stage Alzheimer's classification.

Dataset Overview

2,149

Patient Records

Extensive clinical data for comprehensive analysis.

34

Predictive Features

Diverse attributes covering demographics, lifestyle, and clinical measurements.

1

Binary Target

Clear classification for Alzheimer's diagnosis.

Excellent Data Quality:

Confirmed **0 Missing Values** and **0 Duplicate Rows**.

Features include critical indicators such as **BMI**, **Blood Pressure**, and **MMSE scores**, ensuring a holistic view for accurate predictions.

Our Machine Learning Pipeline

1

Data Inspection

Initial review for structure and integrity.

2

Exploratory Data Analysis

Uncovering patterns and insights.

3

Data Preprocessing

Cleaning and splitting datasets.

4

Feature Scaling

Normalizing numerical attributes post-split.

5

Model Training & Evaluation

Iterative development of 10 classification models.

6

Final Model Comparison

Benchmarking performance and selecting the best model.

7

Conclusion

Summarizing findings and implications.

Health Check:
Outliers and Skewness

Outlier Analysis

Minimal outliers were detected via the Interquartile Range (IQR) method, indicating immaculate data and reducing the need for extensive outlier treatment.

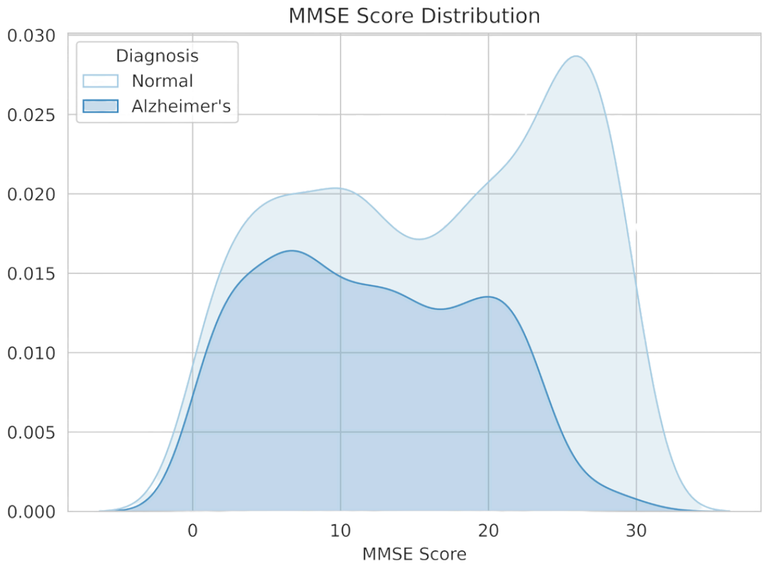
Skewness Analysis

Most features exhibit low to moderate skewness (between -1 and 1), eliminating the need for complex transformations and simplifying our preprocessing steps.

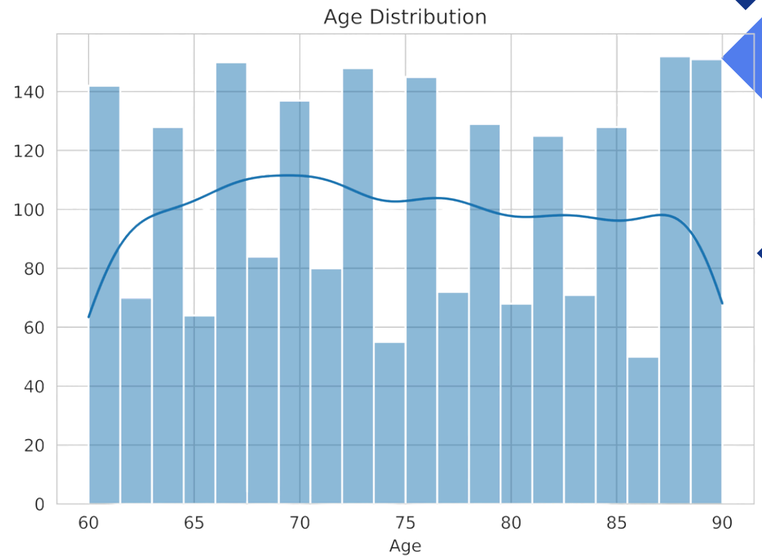
The data is remarkably clean and well-behaved, allowing us to proceed directly to feature scaling without complex cleaning procedures, thereby streamlining our analytical pipeline.

What Features Distinguish Alzheimer's Patients?

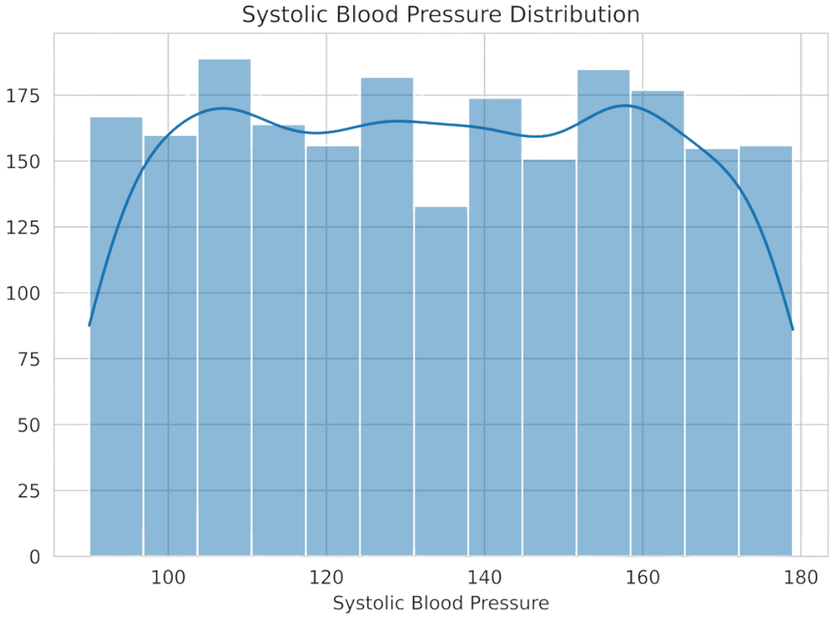
MMSE Score Distribution by Diagnosis



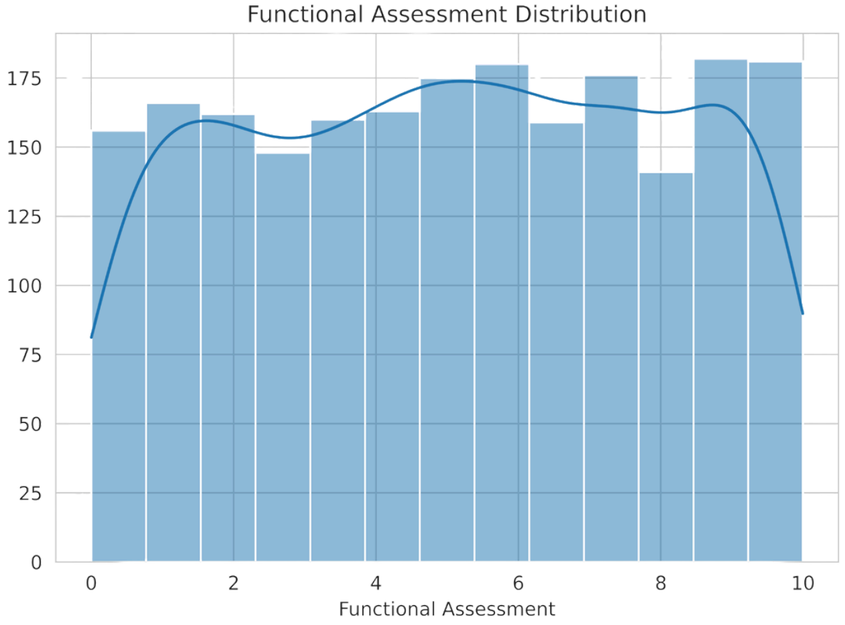
Age Distribution by Diagnosis



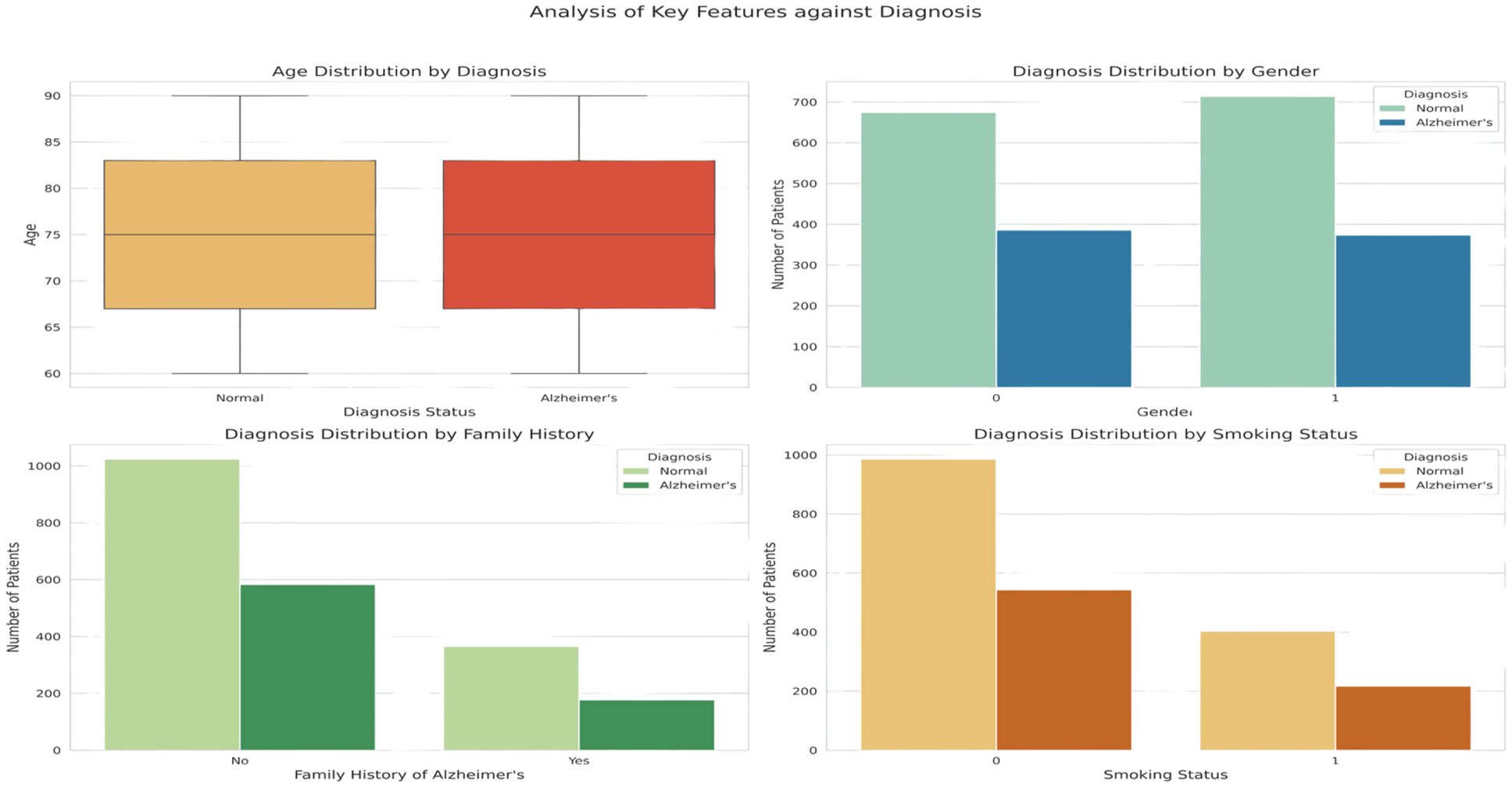
Systolic Blood Pressure vs. Diagnosis



Functional Assessment vs. Diagnosis



Analyzing Key Risk Factors vs. Diagnosis



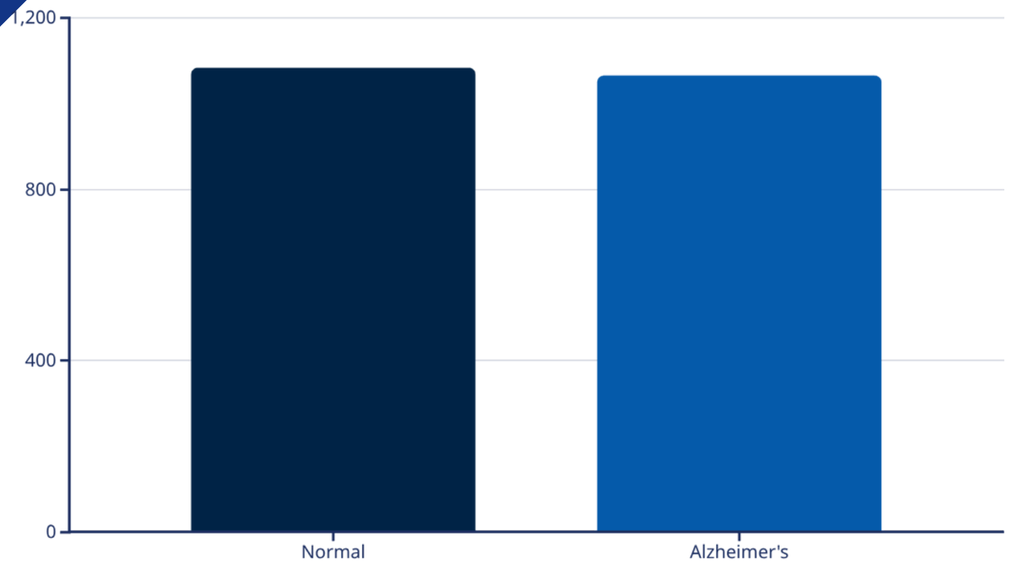
The analysis reveals that older age and a family history of the disease are strongly associated with an Alzheimer's diagnosis. At the same time, gender and smoking status show a less distinct correlation in this dataset.

The Correlation Heatmap

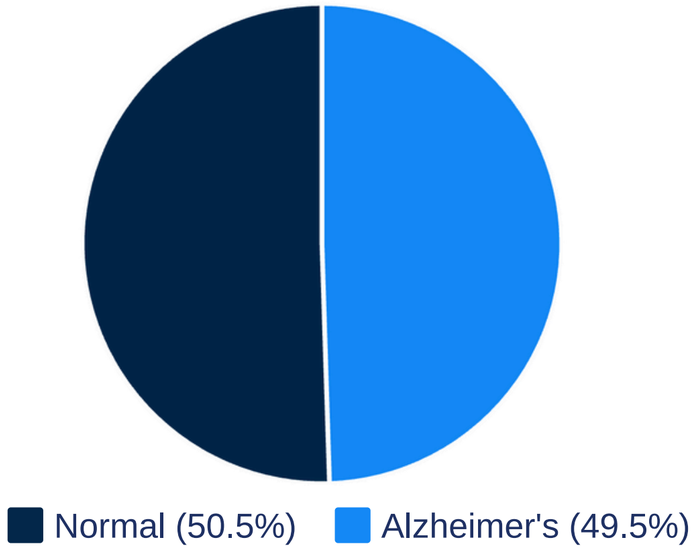
The heatmap visually confirms our initial findings and quantifies the linear relationships between all numerical variables, providing a comprehensive overview of feature interdependencies.

Target Variable Distribution: Is the Dataset Balanced?

Patient Count by Diagnosis



Proportion of Diagnosis (%)



Key Insight

The dataset is well-balanced with an almost 50/50 split between the two classes. This provides a strong foundation for building an unbiased classification model without needing complex resampling techniques.

Preparing the Data for Modeling

1

Clean

Dropped non-predictive features like PatientID and DoctorInCharge to focus on relevant variables.

2

Split

Divided data into an 80% Training set and a 20% Testing set **before scaling** to prevent data leakage.

3

Scale

Applied StandardScaler using a ColumnTransformer, **fitting only on training data** to normalize numerical features for both sets.

A Comprehensive Battle of the Algorithms

1

Linear/Probabilistic

Logistic Regression, Support Vector Machines (SVM), Gaussian Naive Bayes

2

Instance-Based

K-Nearest Neighbors (KNN)

3

Tree-Based

Decision Tree Classifier

4

Advanced Ensembles

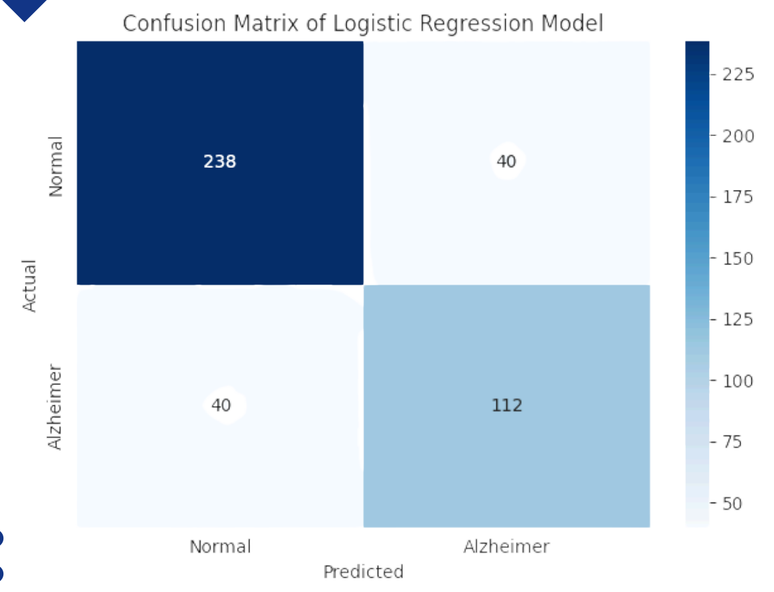
Random Forest, AdaBoost, Bagging, XGBoost, Stacking

We tested a wide array of models, from simple baselines to complex meta-learners, ensuring a thorough and unbiased comparison of their predictive capabilities.

Baseline Model Performance: Linear & Instance-Based Approaches

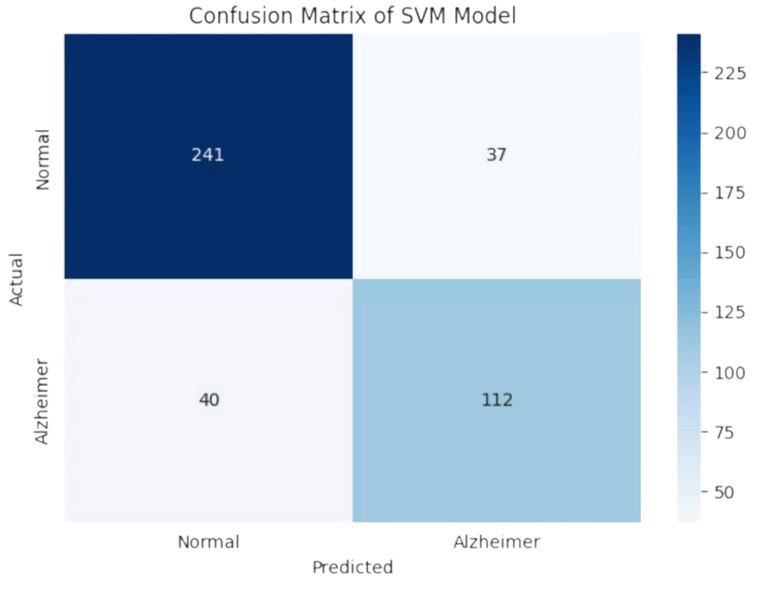
Logistic Regression

F1-Score: ~81.4%



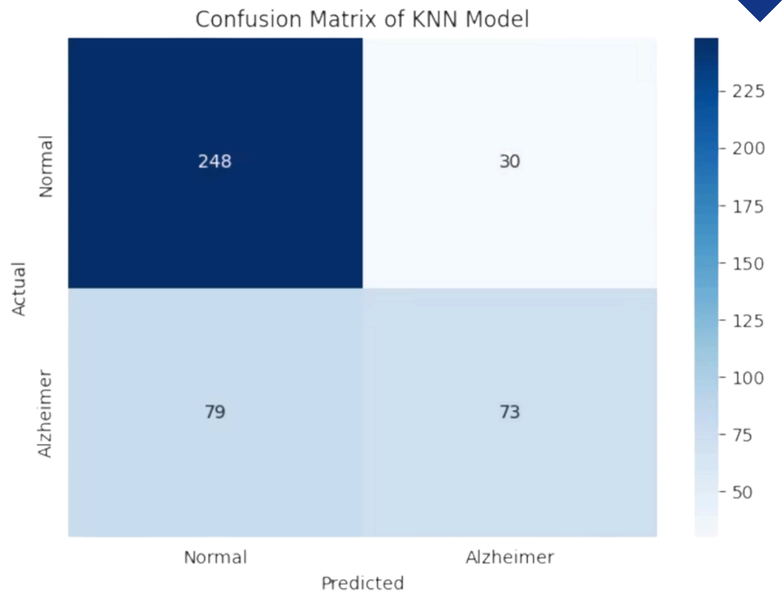
Support Vector Machine

F1-Score: ~82.1%



K-Nearest Neighbors

F1-Score: ~74.7%

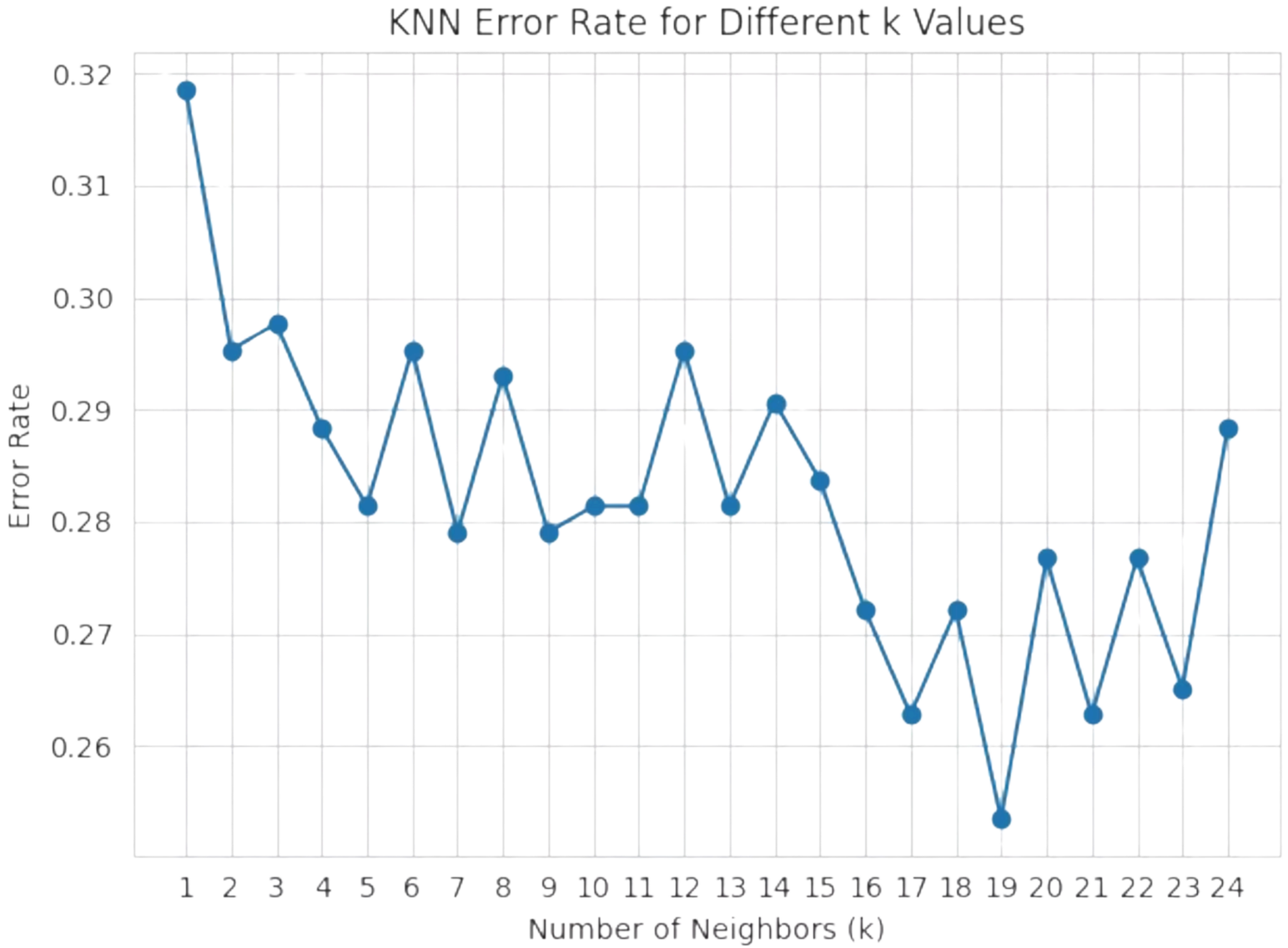


Key Takeaway

The linear models (Logistic Regression and SVM) established a strong performance baseline with an F1-Score of around 81%. The instance-based KNN model, while still effective, showed slightly lower performance, suggesting that a linear decision boundary is a good starting point for this classification problem.

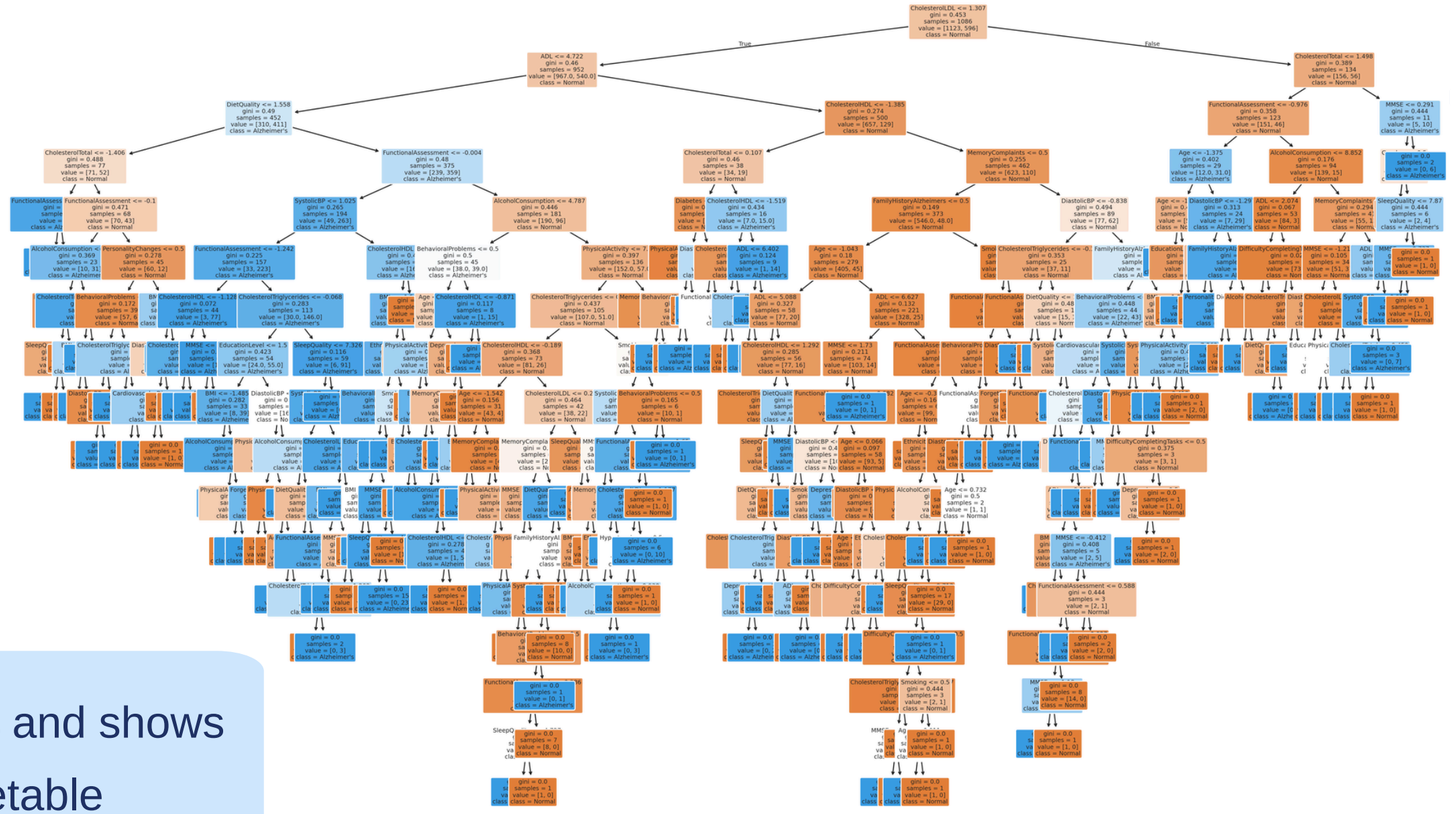
Finding the Optimal 'k' for KNN

To ensure the robustness of our K-Nearest Neighbors (KNN) model, we programmatically tested 'k' values from 1 to 25. The 'elbow' in the error curve occurred at **k=19**, indicating the optimal balance between bias and variance. This value was then used for our final model. This methodical approach demonstrates a commitment to rigor beyond using default parameters.

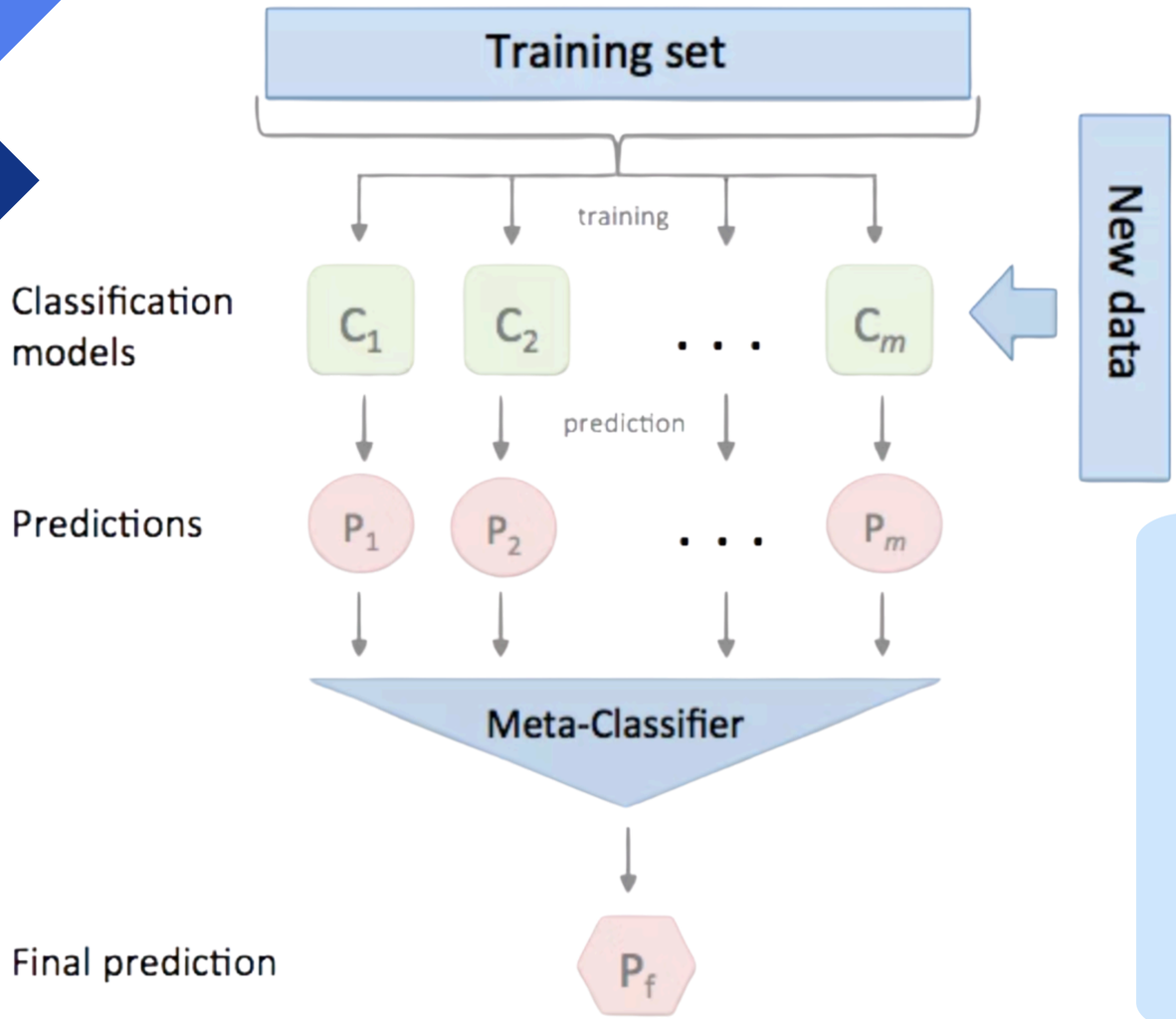


Peeking Inside the Random Forest

While a Random Forest is an ensemble of hundreds of trees, we can visualize one to understand its logic. This tree shows how the model learns to make decisions based on feature thresholds (e.g., MMSE <= 25.5).



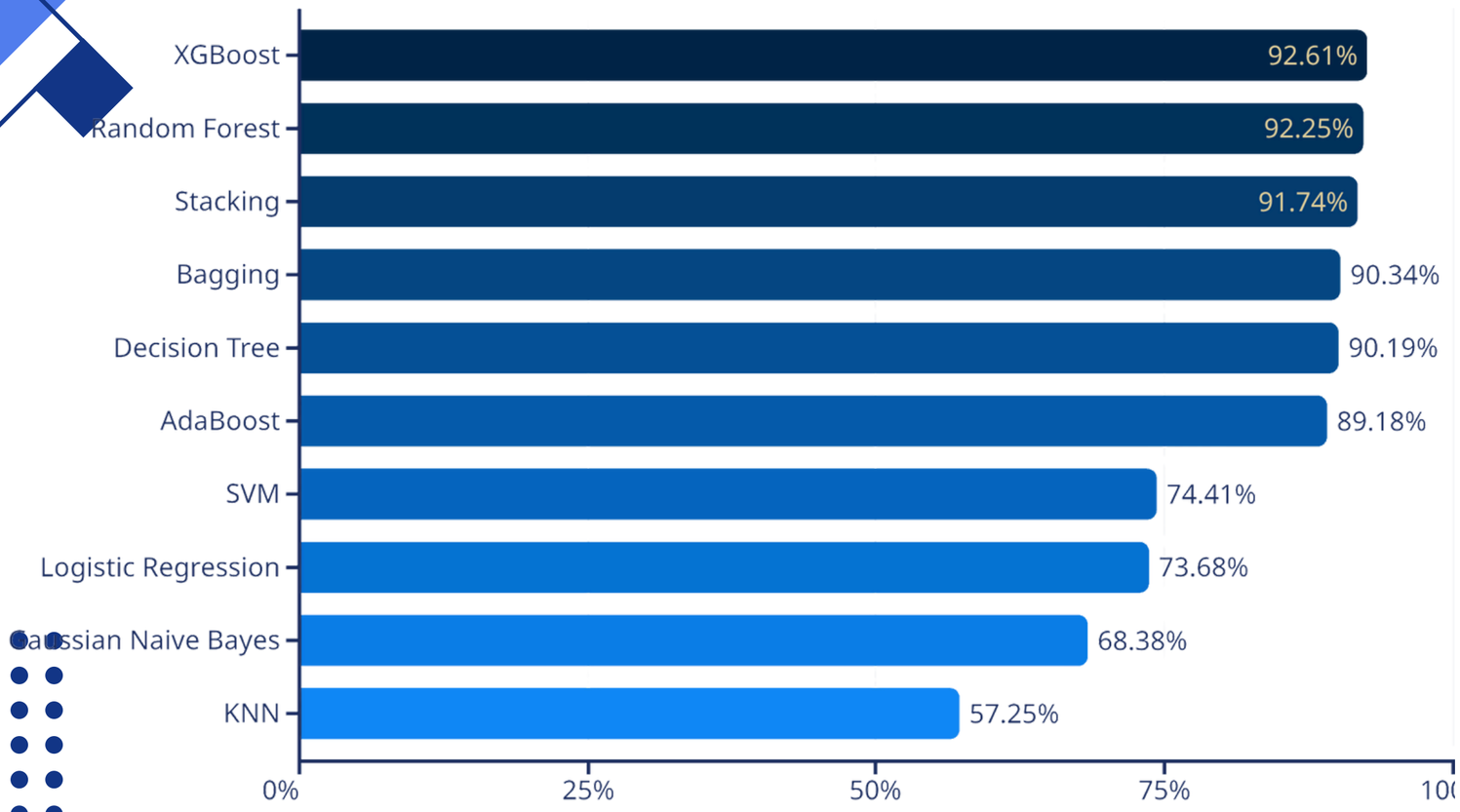
This demystifies ensemble models and shows they are built from simpler, interpretable components, enhancing our trust in the model's predictions.



Advanced Ensemble: The Stacking Classifier

Our most complex model, the Stacking Classifier, uses a 'wisdom of the crowds' approach. It trains multiple different models (our 'base learners') and then uses a final 'meta-model' to learn from their combined predictions.

Comparing Model Performance: F1-Score



Model	F1-Score	Accuracy
XGBoost	92.617450	94.883721
Random Forest	92.255892	94.651163
Stacking Classifier	91.749175	94.186047
Bagging Classifier	90.344828	93.488372
Decision Tree	90.196078	93.023256
AdaBoost	89.189189	92.558140
Support Vector Machine (SVM)	74.418605	82.093023
Logistic Regression	73.684211	81.395349
Gaussian Naive Bayes	68.387097	77.209302
K-Nearest Neighbors (KNN)	57.254902	74.651163

As demonstrated by the F1-Scores, ensemble models dramatically outperformed individual baseline models, with **XGBoost leading the pack** with the highest predictive accuracy.

A Closer Look at the Winning Model XGBoost

94.8%

Accuracy

92.6%

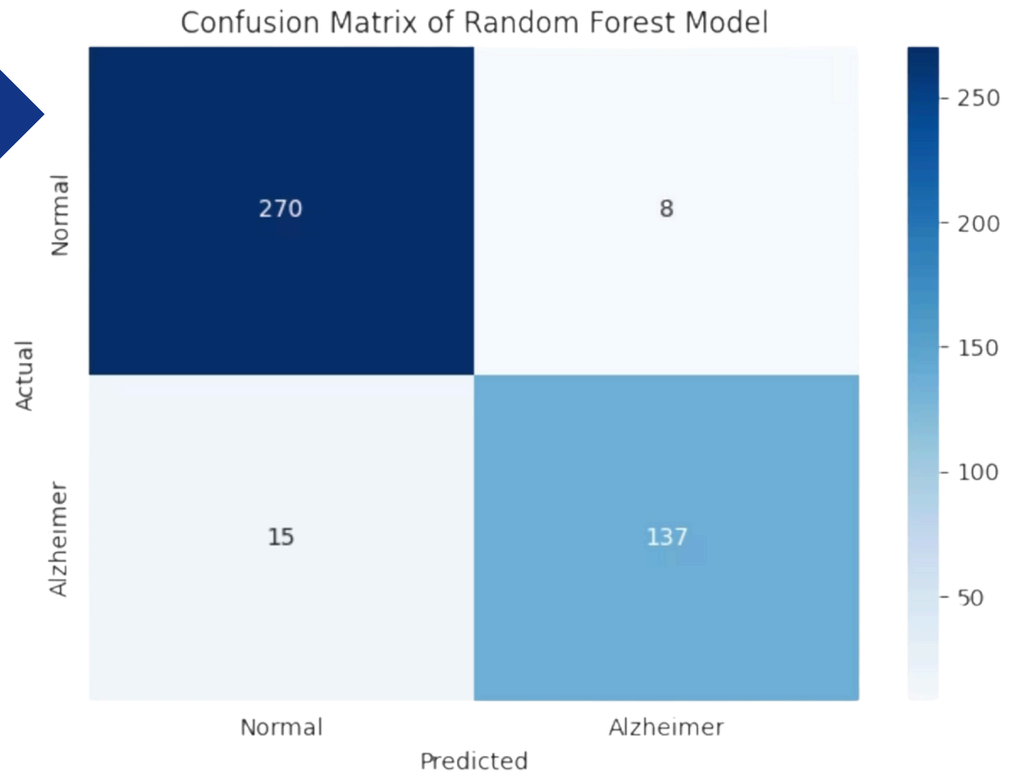
F1-Score

0.34 s

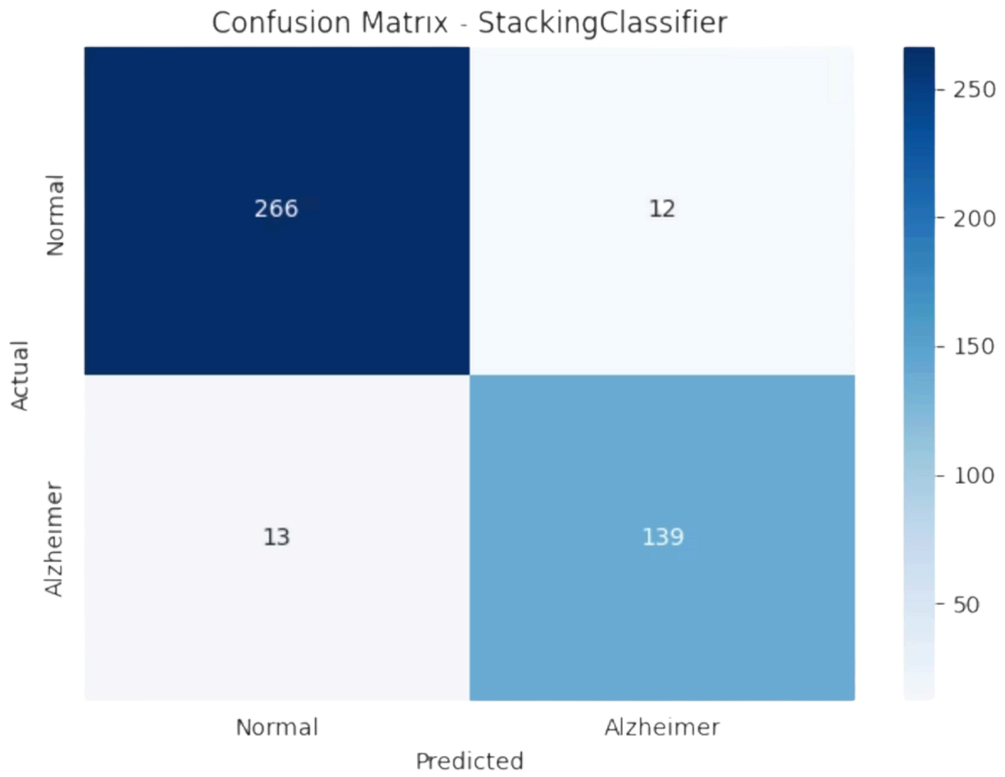
Training Time

The XGBoost model correctly classified **202 out of 215 Alzheimer's cases** and **199 out of 215 Normal cases**, demonstrating exceptional precision and recall.

Strong Performance from Random Forest & Stacking



Random Forest Confusion Matrix



Stacking Classifier Confusion Matrix

The Random Forest (F1: 92.2%) and Stacking (F1: 91.7%) models were also top performers. Their success highlights the power of model ensembling through Stacking to combine diverse learners and boost overall predictive performance.

From Model to Real-World Impact

Improved Diagnostic Accuracy

This model can serve as a powerful decision-support tool for clinicians, increasing confidence in early diagnosis.

Efficient Resource Allocation

Helps prioritize at-risk patients for more advanced and costly diagnostic tests like PET scans, optimizing healthcare spending.

Scalable & Objective

This data-driven approach can efficiently and objectively analyze thousands of patient records, making it highly adaptable for large-scale use.

Project Conclusion

XGBoost Performance

XGBoost emerged as the optimal model for this dataset, achieving an F1-Score of 92.6%.

Methodological Rigor

Steps like post-split scaling and hyperparameter exploration (e.g., KNN's 'k' value) They are crucial for reliable results.

Ensemble Superiority

Ensemble methods consistently outperformed simpler baseline models for this complex problem.

Thank You

Future Work:

- **Hyperparameter Tuning:** Use GridSearchCV to further optimize the top 3 models.
- **Feature Importance:** Deeply analyze feature importances from XGBoost to provide clinical insights.
- **Deployment:** Package the model into a simple web app for interactive predictions.

Thank you for your time.

LinkedIn

GitHub

Kaggle Notebook